УДК 004.932.751

Ю.В. Крак, А.С. Тернов, М.П. Лісняк
Інститут кібернетики ім. В.М. Глушкова НАН України, м. Київ, Україна
yuri.krak@gmail.com, anton.ternov@gmail.com

Структурно-віземний аналіз артикуляції українського мовлення

У статті пропонується підхід до структурно-віземного аналізу візуальної складової мовленнєвого процесу у відеопотоці. Підхід дозволяє отримувати інформацію про кількісну присутність візем з заданого базового набору на кадрі анімації при обчисленні параметрів оптимального стану тривимірної моделі голови людини. Проведені експериментальні дослідження показали можливість використання запропонованої моделі для ідентифікації базових станів губ при артикуляції на тестовій вибірці відеофрагментів 185 слів української мови.

Вступ і постановка задачі

Захоплення та аналіз жестикуляції й виразів обличчя стали важливою частиною різноманітних мультимедійних систем та інформаційних технологій інтелектуалізації комп'ютерних інтерфейсів, невід'ємною складовою яких є системи комп'ютерного синтезу і аналізу візуальної інформації. Такі системи особливо важливі для людей, що мають вади слуху, адже вони компенсують втрату звукового каналу сприйняття інформації зоровим завдяки сприйманню невербальної міміки і жестикуляції та читанню по губах [1]. Навіть люди з нормальним слухом та навиком мовного спілкування підсвідомо використовують інформацію про рух губ і вирази обличчя, що було підтверджено ефектом Мак-Гурка [2]. Хоча візуальний алфавіт і є неповним [3], на практиці він широко використовується сурдоперекладачами жестової мови при перекладі художньої, наукової, юридичної інформації, доповнюючи жестикуляцію в тих випадках, коли досить суттєво правильно передати зміст речення, інформаційного повідомлення, зберігши граматичну структуру речення розмовної мови.

Крім того, розвиток комп'ютерної техніки, виникнення та розвиток новітніх підходів до аналізу, організації, зберігання та подання інформації робить можливим створення інформаційних технологій у сфері моделювання та аналізу комунікаційної жестової мови розробки інтерактивних систем навчання і контролю знання з використанням три вимірних моделей людини [4]. В цьому контексті аналіз міміки емоції і артикуляції при відтворенні жестової мови є досить важливою віхою в розумінні семантики інформації, що передається, і дасть змогу правильно ідентифікувати внутрішній стан та ставлення людини до повідомлення, сприятиме розумінню та більш правильному аналізу сенсу жестикуляції, створенню реалістичного інформаційного каналу зворотного зв'язку.

Одним з перспективних напрямків розробки систем навчання жестовій мові є створення системи навчання правильній артикуляції, основною задачею якої була б можливість контролювати правильність артикуляції губ при промовлянні чи імітації вимови слів української мови, порівнюючи її з еталонною.

Використання тривимірної моделі голови людини для синтезу реалістичної анімації мовленнєвого процесу та аналізу зміни стану губ людини на відео, з одного боку, не зменшує сприймання навчальної інформації [5], а з іншого – підвищує порівняно з відеоматеріалами інтерактивність і програмну гнучкість систем навчання з огляду на

можливість програмної інтеграції в різні мультимедійні комп'ютерні системи. Під інтерактивністю розуміється можливість перегляду процесу анімації артикуляції з різних ракурсів, з різною швидкістю і високою якістю анімації.

В роботах [3], [6-9] для різних мов на неповній тестовій множині візем було отримано результати розпізнавання елементів візуального алфавіту в межах від 20 до 90% при використанні лише візуального каналу передачі інформації. Тому на першому етапі розробки власної системи аналізу прийнятним вважався би результат розпізнавання, який був би подібний до попередніх досліджень.

Проведений аналіз літературних джерел визначив напрямок досліджень і постановку задачі.

Постановка задачі. Необхідно розробити систему для розпізнавання стану моделі на тестовому наборі елементів візуального алфавіту української мови та відтворити цей стан на тривимірній моделі голови людини. Для досягнення поставленої мети виникла необхідність в розв'язанні наступних задач:

– побудувати модель аналізу артикуляції;

– вибрати основні ознаки, що характеризують поточний стан губ і рота;

– розробити алгоритми обчислення положення точок для відображення та анімації високополігональної тривимірної моделі голови людини.

Алгоритм отримання характеристичних точок

Важливу роль у розпізнаванні на основі візуальної інформації відіграють процеси сегментації та виділення губ на зображенні, знаходження на губах певних характерних точок. В дослідженнях, присвячених візуальному мовленню, найчастіше використовуються такі характерні точки: кути губ, точки, які утворюють дугу Купідона, та найнижчі точки [6], [7].

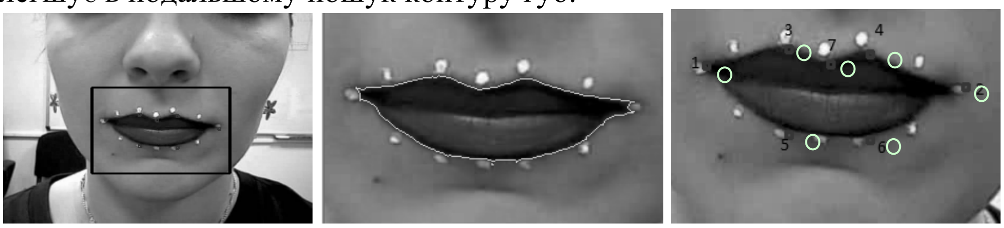
Для попередньої обробки візуальної відеоінформації використовувалась обгортка EmguCV [10] бібліотеки алгоритмів комп'ютерного зору, обробки зображень і чисельних алгоритмів загального призначення з відкритим кодом OpenCV (Open Source Computer Vision Library), яка має достатньо широку функціональність для швидкої цифрової обробки відеозображень.

До кожного кадру застосовується перетворення зображення з кольорового в зображення у градаціях сірого.

Алгоритм обробки першого кадру відео відрізняється від обробки решти кадрів.

За допомогою раніше натренованої на визначення ділянки рота системи каскадної класифікації за Хааром, що поставляється разом з бібліотекою комп'ютерного зору OpenCV, знаходиться ділянка губ. Приклад визначеної цим методом ділянки наведено на рис. 1 а).

Таким чином зменшується розмірність вхідної інформації для подальшої обробки зображення, відокремлюється ділянка з губами від фону та інших частин обличчя, що полегшує в подальшому пошук контуру губ.



а) б) в)

Рисунок 1 – Визначення ділянки губ, контуру і характеристичних точок на кадрі

Наступним кроком алгоритму є знаходження контурів губ. Для цього застосовується метод активних контурів [6], що часто називають «snake». На рис. 1 б) показано результат роботи даного алгоритму на тестовому зображенні.

Після визначення контуру губ проводиться локалізація характерних точок на губах. В дослідженнях обчислюється положення семи точок (рис. 1 в): кути губ (точки 1 і 2), точки, які утворюють дугу Купідона (точки 3, 4 і 7), та точки нижньої губи (точки 5 та 6). Для їх визначення застосовується вертикальне та горизонтальне проектування. Мінімум і максимум горизонтальної проекції контуру дадуть точки, що відповідають кутам губ, тобто точки 1 та 2. Для знаходження точки 3 шукається максимум вертикальної проекції точок контуру. Для точки 4 береться максимум вертикальної проекції точок, що лежать справа від центру ділянки губ. Для знаходження точок 5 та 6 використовується проекція точок 3 та 4 на нижню частину контуру. Точка 7 знаходиться як мінімум вертикальної проекції точок, що лежать між точками 3 та 4.

На всіх наступних кадрах для визначення описаної множини характерних точок було використано алгоритм Лукаса-Канаде слідкування за точками на відео [11]. На вхід цьому алгоритму подається попереднє та поточне зображення в градаціях сірого та координати точок на попередньому зображенні. На виході алгоритму отримаємо координати точок на поточному зображенні. Алгоритм добре працює для точок, що мають деякі особливості на зображенні в своєму околі, наприклад, різні кути чи заломлення, оскільки він порівнює околи точок на різних кадрах. На практиці для нижньої губи можливо виникнення ситуації випинання, при якій суттєво змінюється візуальний портрет губ і для нижніх точок 5 та 6 цей алгоритм починає некоректно працювати. Тому обчислюючи положення характеристичних точок 1, 2, 3, 4, 7 за алгоритмом Лукаса-Канаде, будемо проводити корегування розташування для точок нижнього контуру губ (точки 5 та 6). Нове положення визначатиметься як проекції точок 3 і 4 на контур губ, отриманий за допомогою активних контурів (рис. 2).

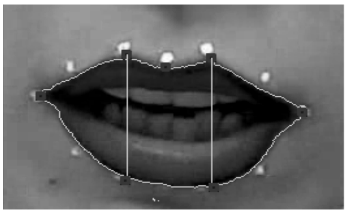


Рисунок 2 – Корекції положення нижніх точок 5 і 6

Загальна схема алгоритму обчислення характеристичних точок наведена на рис. 3.

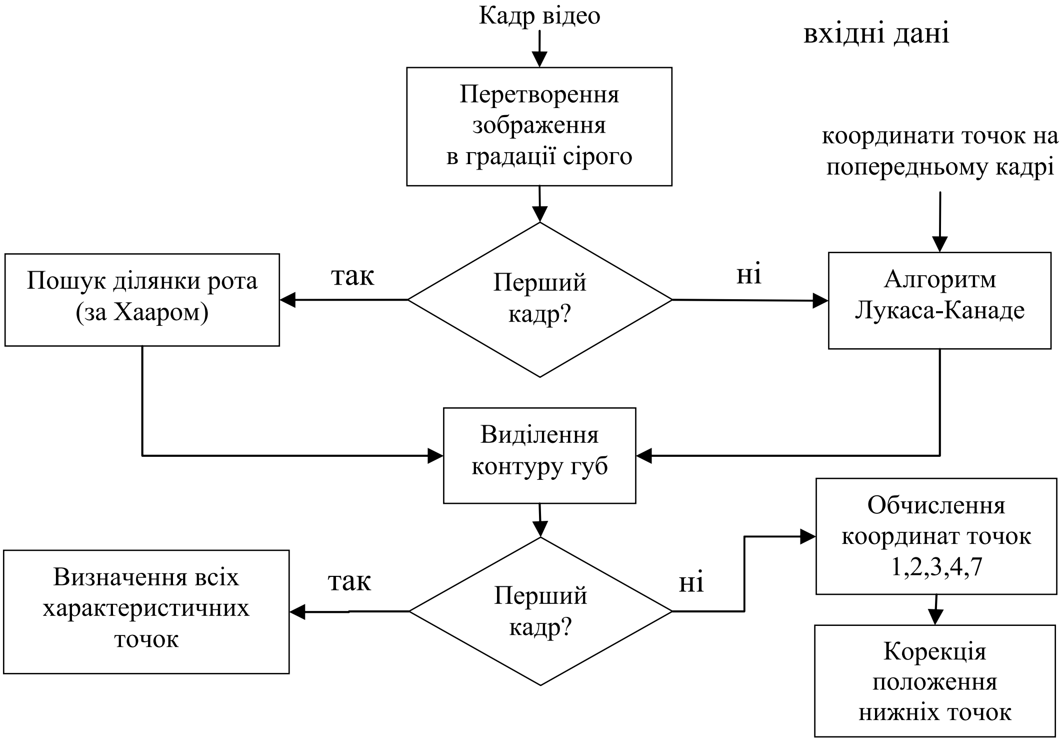


Рисунок 3 – Схема алгоритму отримання характеристичних точок на кадрі анімації

Обчислення характеристичних ознак стану губ. Лінійна модель артикуляції

Вибір візуальних дескрипторів має враховувати особливості артикуляції всіх візем і максимально відображати характерні рухи губ. В роботі розглянуто кілька візуальних дескрипторів: горизонтальна відносна відстань між кутами губ (відстань між точками 1 та 2), зміна якої є характерною для візем «ІИ», «Е», «СЗЦ» та інші; різні вертикальні відносні відстані (відстань між точками 3 і 5, 4 і 6, 3 і 6, 5 і 5) для «А», «ПБМ»; міра округлості губ (площа шестикутника 1-3-4-2-6-5, розділена на площу круга, що побудований на діаметрі 1-2) для «О» та «У». Величини наведених дескрипторів обчислюються за координатами знайдених раніше характеристичних точок (рис. 6).

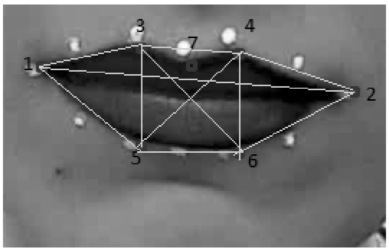


Рисунок 4 – Візуальні дескриптори для семи характеристичних точок

Варіації та зміна вертикальних відстаней між верхньою і нижньою губою дає характеристику ступеня відкритості рота, що дуже важливо для звуків, в яких основна артикуляція відбувається через притиснення губ або широке розкриття губ, наприклад для звуків п, б, а («ПБМ», «А»).

Варіації та зміна горизонтальної відстані дає характеристику ступеня розтягнення кутиків губ, що є важливим для звуків, які утворюються за допомогою розтягнення, наприклад е, и (віземи «ІИ», «Е», «СЗЦ», «КГ»).

Міра округлості дає можливість проаналізувати, наскільки схожий на круг даний контур губ, що дає можливість отримати характеристику для лабіалізованих звуків о, у (віземи «О», «У»).

Вектор ознак матиме наступний вигляд $v = [w, h_1, h_2, c]$, де:

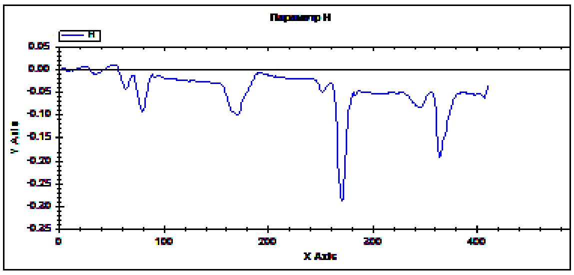
$$h_i = \frac{h_{i_{current}} - h_{i_{default}}}{h_{i_{default}}} \quad i = 1,2 \tag{1}$$

$$w = \frac{w_{current} - w_{default}}{w_{default}} \tag{2}$$

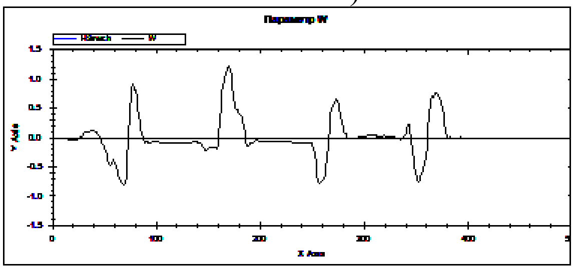
$$c = \frac{S_p}{S_c}, \tag{3}$$

де $h_{current}$ – відстань між точками 3 і 5 на поточному кадрі, $h_{default}$ – відповідна їй відстань в стані спокою, $w_{current}$ – відстань між точками 1 і 2 в даному кадрі, $w_{default}$ – відстань між ними в стані спокою, S_p – площа багатокутника (1,3,4,2,6,5), S_c – площа круга, побудованого на діаметрі, що утворюється точками 1 і 2.

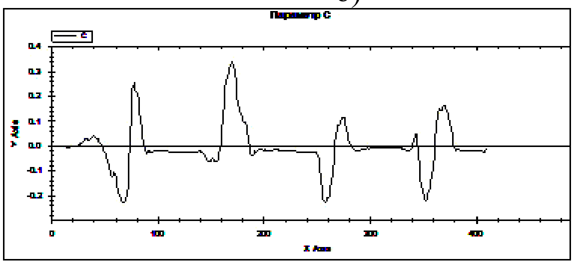
На рис. 5 показана динаміка зміни характеристичних параметрів для слів «бак», «арка», «бук», «борщ».



а)



б)



в)

Рисунок 5 – Динаміка зміни характеристичних параметрів для слів «бак», «арка», «бук», «борщ»: а) параметр h, б) параметр w, в) параметр c

Лінійна модель артикуляції. У даній роботі розглядаються тільки переходи «голосна-приголосна» або, навпаки, «приголосна-голосна». Нехай $w = (x_1, \dots, x_7)$ – вектор характеристичних точок, що відповідають певному положенню губ, певній віземі, де $x_i = (x,\ y,\ z)$ – координати точки в просторі. Ω_1 – множина віздем, що відповідають голосним фонемам, а Ω_2 – множина віздем, що відповідають приголосним фонемам. Вектор характеристичних точок, що відповідає стану спокою, включимо в обидві множини $w_{default} \in \Omega_1$ і $w_{default} \in \Omega_2$.

Введемо відображення F, яке буде задавати перетворення вектора характеристичних точок у вектор параметрів для характеристики віземи.

$$F(w, w_{default}) = \begin{pmatrix} \dfrac{\|x_2 - x_1\| - \|x_{2d} - x_{1d}\|}{\|x_{2d} - x_{1d}\|} \\ \dfrac{\|x_3 - x_5\| - \|x_{3d} - x_{5d}\|}{\|x_{4d} - x_{5d}\|} \\ \dfrac{\|x_4 - x_6\| - \|x_{4d} - x_{6d}\|}{\|x_{4d} - x_{6d}\|} \\ \dfrac{S_p}{S_c} \end{pmatrix}, \qquad (4)$$

де S_p – площа багатокутника, побудованого на точках 1, 3, 4, 2, 6, 5, а S_c – площа круга, побудованого на точках 1, 2 як на діаметрі.

Оскільки в українській мові поточний стан артикуляції залежить від промовляння двох фонем [12], будемо їх шукати як деяку комбінацію двох базисних станів, що є найближчою до даного стану.

Будемо використовувати два підходи. Перший підхід полягає в тому, щоб шукати параметр α, базуючись на векторі характеристичних точок кожної базисної віземи.

$$\alpha = \arg \min_{\substack{w_j \in \Omega_1 \\ w_j \in \Omega_2 \\ w_i \neq w_j \\ \alpha \in [o,1]}} \rho(F(\alpha w_i + (1-\alpha) w_j), F(w_{curr})) . \qquad (5)$$

Назвемо (5) лінійною моделлю артикуляції (ЛМА).

У другому підході потрібно спочатку знайти вектори параметрів, що характеризують віземи для кожного базисного стану, а вже після цього на їх основі шукати параметр α

$$\begin{aligned} v_i &= F(w_i),\ w_i \in \Omega_1 \\ v_j &= F(w_j),\ w_j \in \Omega_2 \end{aligned} . \qquad (6)$$

Тоді формула (5) буде мати вигляд:

$$\alpha = \arg \min_{\substack{w_j \in \Omega_1 \\ w_j \in \Omega_2 \\ w_i \neq w_j \\ \alpha \in [o,1]}} \rho(\alpha v_i + (1-\alpha) v_j, F(w_{curr})) . \qquad (7)$$

Для розв'язання оптимізаційної задачі (5) та (7) можна використовувати різні методи і підходи, наприклад метод динамічного програмування, але враховуючи низьку розмірність задачі (максимально можливих комбінацій лише 255) та відсутність потреби у великій точності (з огляду на похибки вимірювання не більші ніж до другого знаку після коми), розв'язок можна отримати повним перебором.

Нейронні мережі. Для використання розпізнавання на основі штучних нейронних мереж використовувалась бібліотека роботи з штучними нейронними мережами Encog [12]. В дослідженнях розглядались дві прямопрохідні багатошарові нейронні мережі з однаковою внутрішньою структурою. Вони містили три нейрони у вхідному шарі, двадцять п'ять нейронів у прихованому та вісім нейронів, що характеризують віземи, у вихідному шарі.

У прихованому шарі використовувалась сигмоїдна активаційна функція:

$$x_j^p = \sum_i w_{ji} y_i^p \qquad y_j^p = \sigma(x_j^p) = \frac{1}{1+e^{-x_j^p}} \qquad (8)$$

x_j^p – вхід на нейрон j, y_j^p – вихід з нейрону j, w_{ij} – вага між нейронами i та j, t_j^p – цільовий вихід з j-го нейрона при навчальному шаблоні *p*.

У вихідному шарі використовувалась функція, що масштабує значення так, щоб в сумі вони дорівнювали одиниці.

Перша мережа була натренована на тестових зображеннях (кадрах з відеозразків бази слів, що відповідають певній віземі). На цих кадрах запускався алгоритм локалізації характерних точок та алгоритм обчислення характеристичних параметрів для стану губ. Отримані характеристичні параметри використовувались як тренувальна вибірка для першої нейронної мережі (НМ1).

Для другої нейронної мережі (НМ2) визначались характеристичні параметри стану губ для тривимірної моделі для кожної віземи з не більш ніж п'ятипроцентним збуренням. Тобто характеристичні параметри визначались для морфа $w = \alpha \cdot w_i + (1-\alpha) \cdot w_j$, де $\alpha \in [0.95,1]$.

Анімація на тривимірній моделі голови людини

Для відтворення та анімації на основі лінійної моделі використовувались морфемна комп'ютерна анімація та алгоритми морфінгу [13]. Для оптимізації обчислення використовувались технології обчислення положення точок за допомогою графічного процесора та програмних шейдерів.

Шейдер (англ. Shader) – це програма для одного із ступенів графічного конвеєра, що використовується в тривимірній графіці для визначення остаточних параметрів об'єкта чи зображення. Вона може містити у собі довільної складності описи поглинання та розсіювання світла, накладення текстури, віддзеркалення і заломлення, затінення, зміщення поверхні і ефекти постобробки [14]. Програмовані шейдери гнучкі та ефективні. Складні на вигляд поверхні можуть бути візуалізовані за допомогою простих геометричних форм. Обчислення за допомогою шейдерів виконуються приблизно в 20 раз швидше від обчислень на центральному процесорі [15]. Для швидкого розрахунку морфемної анімації використовуються вершинні шейдери. Для цього у відеопам'ять заноситься меш точок голови, текстур та нормалей для всіх базисних станів. Це статична інформація, що записується лише раз і там зберігається. Після цього у вершинний шейдер на кожному кроці передається параметр α та індекси базових візем, на основі яких обчислюється значення кінцевого положення кожної точки.

Ефективність цієї методики полягає в тому, що графічний процесор, на якому виконуються шейдерні підпрограми, оптимізовано для багатопотокового, паралельного обчислення положення великої кількості точок. Крім того, при такому підході основний масив даних знаходиться у відеопам'яті, тому не витрачається час на копіювання даних з неї до оперативної пам'яті комп'ютера і навпаки для проведення обчислень та подальшого відображення кінцевої стану тривимірної моделі.

Результати експериментальних досліджень

Для проведення експериментальних досліджень на основі вищеописаних алгоритмів та методів було розроблено оригінальне програмне забезпечення (рис. 6) з наступною функціональністю:

1) виділення ділянки обличчя з кадру зображення;
2) виділення контуру губ та локалізація характеристичних точок;
3) визначення вектора характеристичних параметрів;
4) побудова графіків зміни характеристичних параметрів, знешумлених за допомогою вейвлетів Добеші [16];
5) побудова штучних нейронних мереж для розпізнавання;
6) розпізнавання конкретного статичного стану губ на кадрі відео за допомогою лінійної моделі артикуляції і за допомогою штучних нейронних мереж.

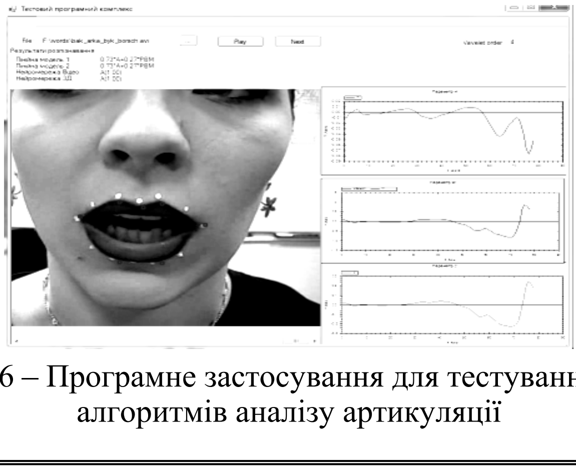


Рисунок 6 – Програмне застосування для тестування роботи алгоритмів аналізу артикуляції

Для коректної роботи програми на зображення чи кадр відео накладаються такі обмеження:

– обличчя людини на зображенні повинно займати не менш ніж 50% площі фотографії чи кадру;

– обличчя людини нахилене не більше ніж під кутом 10°, щоб кутики губ були по вертикальній осі нижче, ніж точки дуги Купідона. Щоб позбутися цього обмеження, потрібно вводити поправку на орієнтацію обличчя;

– обличчя людини повинне бути освітлене досить рівномірно та колір губ повинен суттєво відрізнятися від кольору шкіри;

– перший кадр відео відповідає стану спокою губ.

Окремо було розроблено програмне забезпечення для відтворення базових візем та їх комбінацій на тривимірній моделі голови людини. Обчислення положення точок для комбінацій візем проводиться за допомогою шейдерів.

Для задач реалістичного відтворення міміки та артикуляції використовується високополігональна модель голови людини. Ця модель отримана за допомогою програмного пакета Poser. Модель складається з 20 000 точок, що утворюють 60 000 трикутників та текстур голови, очей, зубів [13].

Також було реалізовано міжпроцесну комунікацію між цими двома програмами, для відтворення на тривимірній голові людини розпізнаного стану. Результат роботи програмного комплексу показано на рис. 7.

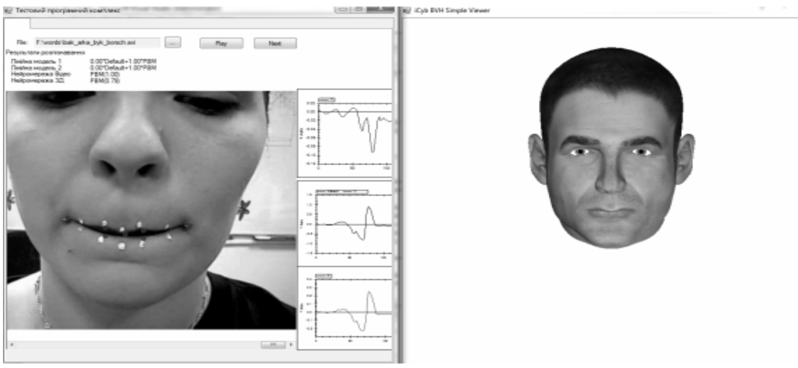


Рисунок 7 – Вікна програмного комплексу аналізу і синтезу артикуляційного процесу

Обчислювальний експеримент проводився на відеобазі віземних переходів для української мови. Елементи відеобази були отримані з множини фонематично записаних слів-зразків, яка покриває множину всіх комбінацій візем української мови CV(VC)-типу (комбінація «приголосний-голосний», «голосний-приголосний»). При визначенні множини слів-зразків використовувались дослідження в галузі фонетики і фонології української мови [17], [18], які визначають основні артикуляційні моменти відповідно до фонемно-алофонного подання морфем і слів української мови. Загальна кількість слів-зразків 185.

Для спрощення задачі автоматичного пошуку характеристичних точок контур губ був наведений чорним кольором.

Після цього були пронумеровані кадри, на яких присутній явний вигляд певної віземи, сформована навчальна та тестова вибірки. У тестовій вибірці міститься по 20 зображень кожної віземи.

Якщо не враховувати визначення стану спокою, який завжди коректно ідентифікувався, результат розпізнавання для нейронної мережі, що була натренована на тестовій вибірці з цього ж відео, у найгіршому випадку становив 35% для фонеми Е, а найкращий – 85% для О та У. В табл. 1 наведено результати чисельного експерименту для всіх тестових візем.

Таблиця 1 – Матриця прийнятих рішень на тестовому наборі НМ1

	А	ПБМ	О	У	Е	Спокій	ІИ	СЗЦ	КГ	Невиз.
А	**14**	0	0	0	3	0	0	0	0	3
ПБМ	0	**16**	0	0	4	0	0	0	0	0
О	0	0	**17**	3	0	0	0	0	0	0
У	0	0	3	**17**	0	0	0	0	0	0
Е	2	0	0	2	**7**	0	2	2	0	5
Спокій	0	0	0	0	0	**20**	0	0	0	0
ІИ	0	0	0	2	4	0	**12**	0	0	2
СЗЦ	0	0	0	0	0	0	12	**8**	0	0
КГ	0	0	0	0	0	0	4	8	**8**	0

В табл. 2 наведено результат експерименту для нейронної мережі, що була натренована на базових морфах тривимірної моделі голови людини. Результати виявились гіршими порівняно з попередньою. Найкращим був результат розпізнавання стану губ для віземи ПБМ – 80%, найгірший – для віземи СЗЦ – 10%. Таке погіршення обумовлено тим, що морфеми візем були раніше синтезовані на основі відеозаписів процесу артикуляції іншої людини і відмінності артикуляції виявились суттєвими для розпізнавання візем «СЗЦ», «КГ», «Е». Крім того погіршення розрізнення візем «О» та «У» свідчить про більшу подібність між собою відповідних їм морфем, ніж відповідних візуальних портретів на кадрах тестової вибірки.

Таблиця 2 – Матриця прийнятих рішень на тестовому наборі для НМ2

	А	ПБМ	О	У	Е	Спокій	ІИ	СЗЦ	КГ	Невиз.
А	**10**	0	0	0	0	0	0	2	6	2
ПБМ	0	**16**	0	0	0	4	0	0	0	0
О	8	0	**8**	4	0	0	0	0	0	0
У	4	1	9	**6**	0	0	0	0	0	0
Е	2	2	0	0	**4**	0	0	10	0	2
Спокій	0	0	0	0	0	**20**	0	0	0	0
ІИ	0	0	0	2	4	0	**2**	10	0	2
СЗЦ	0	0	0	2	6	0	2	**4**	2	4
КГ	0	0	0	4	6	0	0	0	**4**	4

Таблиця 3 – Матриця прийнятих рішень на тестовому наборі для ЛМА

	А	ПБМ	О	У	Е	Спокій	ІИ	СЗЦ	КГ	Невиз.
А	**16**	0	4	0	0	0	0	0	0	0
ПБМ	0	**17**	0	0	0	3	0	0	0	0
О	5	0	**8**	0	0	4	0	0	0	3
У	5	1	5	**4**	0	0	0	0	0	3
Е	4	0	0	0	**7**	0	2	0	0	5
Спокій	0	0	0	0	0	**20**	0	0	0	0
ІИ	0	0	0	0	6	0	**4**	0	2	8
СЗЦ	0	0	0	0	2	3	2	**3**	0	10
КГ	0	0	0	0	3	2	2	4	**0**	9

В табл. 3 наведено результати, отримані при застосуванні лінійної моделі артикуляції (5,6). Розпізнавання візем тестової вибірки коливалось в межах 10 – 85%. Якщо не брати до уваги, що візема «КГ» взагалі не ідентифікувалась. Результати

розпізнавання візем голосних фонем «о», «у», «е», які помилково ідентифікувались як візема «А», свідчать про переважний вплив на цільову функцію (5) параметра, пов'язаного з відкриттям рота і для подальших досліджень слід буде вводити вагові коефіцієнти впливу параметрів залежно від типу віземи або враховувати додаткову інформацію про артикуляцію для визначення додаткових характеристичних параметрів. Для моделі (7) результати виявились подібними до моделі (5). Різниця була несуттєвою, тому в роботі ці дані не наводяться.

Отже, найкраща якість розпізнавання візем тестової вибірки була отримана за допомогою нейронних мереж, яка була натренована на кадрах відеопроцесу артикуляції тієї самої людини. Хоча лінійна модель артикуляції на основі базових станів візуальних портретів на даному етапі досліджень в декількох моментах працює гірше, вона все одно продемонструвала непогані результати, відповідно до очікуваних. Тому, з огляду на можливість врахування ідентифікації віземних переходів, лінійна модель артикуляції є більш перспективною для аналізу динаміки артикуляції і буде в подальшому адаптована для врахування інформації з попередніх кадрів.

Висновки

Результати експерименту підтверджують працездатність даного підходу, особливо для фонем, для яких характерне вертикальна зміна конфігурації рота. Також було отримано, що візуальні портрети фонем **о** і **у** є досить схожими і вони можуть бути сплутані одна з одною. Для кращої ідентифікації цих візем, з огляду на особливості їх артикуляції, потрібно враховувати дані зображення обличчя в профіль.

Подальші дослідження будуть направлені на удосконалення лінійної моделі аналізу артикуляції за рахунок введення вагових коефіцієнтів для параметрів моделі залежно від типу віземи та врахування динаміки зміни стану губ, використовуючи інформацію з попередніх кадрів. Це сприятиме покращенню якості розпізнавання і надасть можливість до побудови системи розпізнавання шаблонів слів української мови з наступною прив'язкою до особливостей артикуляції при відтворенні жестової мови.

Література

1. Ouni S. Visual Contribution to Speech perception / S. Ouni, M. Cohen, D. Massaro // EURASIP Journal on Audio, Speech and Music Processing. – 2007. – P. 1-12.
2. McGurk H. Hearing lips and seeing voices / H. McGurk, J. MacDonald // Nature. – 1976. – Vol. 264. – P. 746-768.
3. Крак Ю.В. Информационная технология для автоматического чтения по губам украинской речи / Ю.В. Крак, О.В. Бармак, А.С. Тернов // Комп'ютерна математика. – 2009. – № 1. – С. 86-95.
4. Інформаційна технологія для моделювання української мови жестів / Ю.Г. Кривонос, Ю.В. Крак, О.В. Бармак [та ін.] // Штучний інтелект. – 2009. – № 3. – С. 186-197.
5. Beskow J. The Teleface project – disability, feasibility and intelligibility [Електронний ресурс] / J. Beskow, M. Dahlquist, B.Granström et al. – Режим доступу : http://www.speech.kth.se/~beskow/papers/fon97teleface.pdf.
6. Werda S. Lip localization and viseme classification for Visual Speech Recognition / S. Werda, W. Mahdi, Abdelmajid Ben Hamadou // International Journal of Computer & Information Science. – 2007. – C. 62-75.
7. Sadat V. Sadeghit Vowel recognition using Neural Networks / Sadat V. Sadeghit, K. Yaghmaiett // IJCSNS International Journal of Computer Science and Network Security. – 2006. – P. 154-158.
8. Давидов М.В. Алгоритм визначення форми губ під час артикуляції для української жестової мови / М.В. Давидов, Ю.В. Нікольський, С.М. Тиханський // Інформаційні системи та мережі. – Львів : Видавництво Національного університету «Львівська політехніка», 2010. – № 673. – С. 267-273.
9. Мурыгин К.В. Концепция системы распознавания речи на основе чтения по губам / К.В. Мурыгин // Штучний інтелект. – 2009. – № 2. – С. 116-123.
10. Електронний ресурс EmguCV [Електронний ресурс]. – Режим доступу : http://www.emgu.com/wiki/index.php/Main_Page.

11. Bouguet J. Pyramidal Implementation of the Lucas Kanade feature tracker. Description of the algorithm [Електронний ресурс] / J. Bouguet. – Intel Corporation Microprocessor Research Labs. – Режим доступу : http://robots.stanford.edu/cs223b04/algo_tracking.pdf
12. Електронний ресурс Encog project, Heaton Research [Електронний ресурс]. – Режим доступу : http://www.heatonresearch.com/encog.
13. Кривонос Ю.Г. Синтез візуальної складової зовнішньої артикуляції на обличчі людини з використанням морфів візем для моделювання жестової мови / Ю.Г. Кривонос, Ю.В. Крак, А.С. Тернов // Искусственный интеллект. Интеллектуальные системы (ИИ-2010) : материалы Междунар. науч.-техн. конф. (пос. Кацивели, АР Крым, 20 – 24 сентября 2010 г.). – Донецк : ИПИИ «Наука і освіта», 2010. – С. 291-294.
14. Kessenih J. The OpenGL Shading Language [Електронний ресурс] / J. Kessenih // Inc. Ltd. – 2006. – 87 p. – Режим доступу : http://www.opengl.org/registry/doc/GLSLangSpec.Full.1.20.8.pdf
15. Боресков А.В. Разработка и отладка шейдеров / Боресков А.В. – Санкт-Петербург : БХВ, 2006. – 496 с.
16. Добеши И. Десять лекций по вейвлетам / Добеши И. – Ижевск : НИЦ «Регулярная и хаотическая динамика», 2001. – 464 с.
17. Білодід І.К. Сучасна українська літературна мова. Вступ. Фонетика / Білодід І.К. ; Ін-т мовознавства ім. О.О. Потебні. - К. : Наукова думка, 1969. – 435 с.
18. Тоцька Н.І. Сучасна українська літературна мова. Фонетика, орфоепія, графіка, орфографія / Н.І. Тоцька. – К. : Вища школа, 1981. – 182 с.

Literatura

1. Ouni S. EURASIP Journal on Audio, Speech and Music Processing. 2007. P. 1-12.
2. McGurk H. Nature. 1976. Vol. 264. P. 746-768.
3. Krak Ju.V. Komp'juterna matematyka. № 1. 2009. S. 86-95.
4. Krivonos Ju.G. Shtuchnij intelekt. № 3. 2009. S. 186-197.
5. Beskow J. The Teleface project - disability, feasibility and intelligibility. http://www.speech.kth.se/~beskow/papers/fon97teleface.pdf.
6. Werda S. International Journal of Computer & Information Science. 2007. S. 62-75.
7. Sadat V. IJCSNS International Journal of Computer Science and Network Security. 2006. P. 154-158.
8. Davidov M.V Informacijni systemy ta merezhi. L'viv.: Vydavnyctvo Nacional'nogo universytetu "L'vivs'ka politehnika". № 673. 2010. S. 267-273.
9. Murygin K.V. Shtuchnij intelekt. № 2. 2009. S 116-123
10. EmguCV http://www.emgu.com/wiki/index.php/Main_Page.
11. Bouguet J. Pyramidal Implementation of the Lucas Kanade feature tracker. Description of the algorithm. http://robots.stanford.edu/cs223b04/algo_tracking.pdf
12. Krivonos Ju.G. Sintez vizual'noi skladovoї zovnishn'oi artikuljacii na oblichchi ljudeni z vikoristannjam morfiv vizem dlja modeljuvannja zhestovoi movi D: IPII "Nauka i osvita". 2010. S. 291-294
13. Kessenih J. The OpenGL Shading Language. http://www.opengl.org/registry/doc/GLSLangSpec.Full.1.20.8.pdf
14. Boreskov A.V. Razrabotka i otladka shejderov. Sankt-Peterburg: BHV. 2006. 496 s.
15. Dobeshi I. Desjat' lekcij po vejvletam. Izhevsk: NIC "Reguljarnaja i haoticheskaja dinamika". 2001. 464 s.
16. Encog project, Heaton Research. http://www.heatonresearch.com/encog.
17. Bilodid I.K. Suchasna ukrains'ka literaturna mova. Vstup. Foneteka. K. : Naukova dumka. 1969. 435 s.
18. Toc'ka N.I. Suchasna ukrains'ka literaturna mova. Foneteka, orfoepija, grafika, orfografija. K : Vishha shkola. 1981. 182 s.

Y.V. Krak, A.S. Ternov, M.P. Lisniak

Structural-Viseme analysis of Ukrainian Speech Articulation

An approach to the structural analysis of visemes of visual component of speech process in the video stream is proposed in this paper. The approach allows to compute numeric information about presence of a viseme in an animation frame chosen from the given base set by calculating the optimal parameters of state for three-dimensional model of a human head. Experimental studies have shown the efficiency of using the proposed model to identify the basic states of lip articulation by test video samples with 185 words of the Ukrainian language.

Стаття надійшла до редакції 22.06.2011.